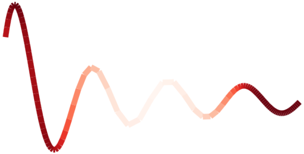
Generalized Energy Based Models

Michael Arbel[1] Liang Zhou[1]
Arthur Gretton[1]

[1]Gatsby Computational Neuroscience Unit
University College London

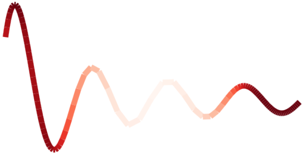
February 22, 2021

Data with low intrinsic dimension: A toy example

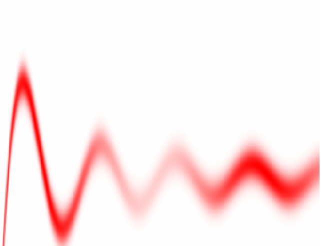


Data

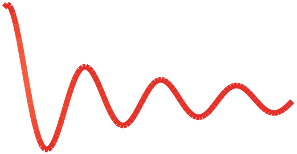
Data with low intrinsic dimension: A toy example



Data

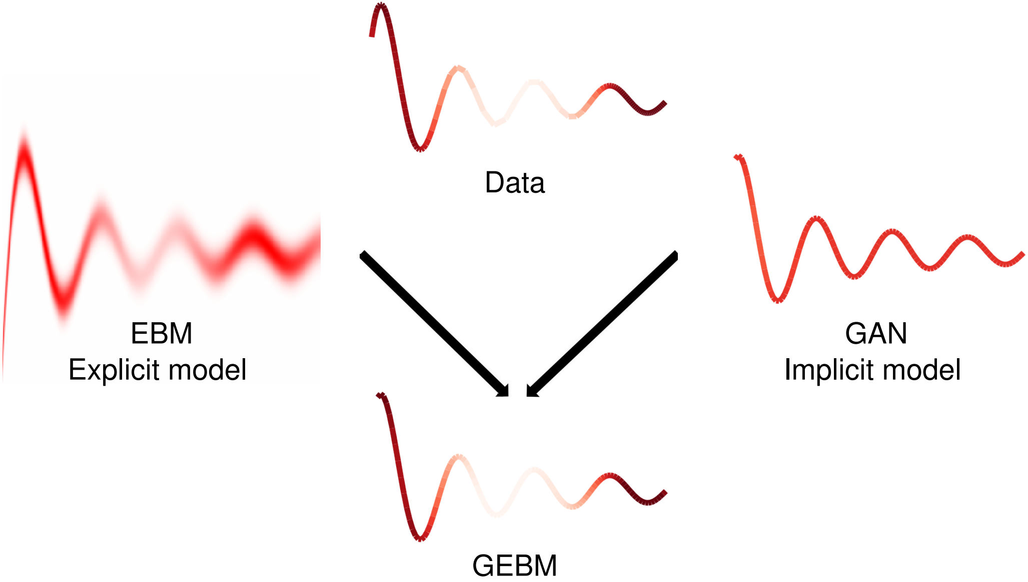


EBM
Explicit model



GAN
Implicit model

Data with low intrinsic dimension: A toy example



Outline

- Data with low intrinsic dimension: The need for new models

Outline

- Data with low intrinsic dimension: The need for new models
- Generalized Energy-Based models: A model with two components
  - The base
  - The energy

Outline

- Data with low intrinsic dimension: The need for new models
- Generalized Energy-Based models: A model with two components
  - The base
  - The energy
- Training GEBMs: a two stages method
  - Learning the energy: Generalized Maximum Likelihood Estimation
  - Learning the base : $KALE$ minimization

Outline

- Data with low intrinsic dimension: The need for new models
- Generalized Energy-Based models: A model with two components
  - The base
  - The energy
- Training GEBMs: a two stages method
  - Learning the energy: Generalized Maximum Likelihood Estimation
  - Learning the base : $KALE$ minimization
- Sampling from GEBMs
  - Latent space MCMC
  - Experimental validation on image datasets.

Outline

- Data with low intrinsic dimension: The need for new models
- Generalized Energy-Based models: A model with two components
  - The base
  - The energy
- Training GEBMs: a two stages method
  - Learning the energy: Generalized Maximum Likelihood Estimation
  - Learning the base : $KALE$ minimization
- Sampling from GEBMs
  - Latent space MCMC
  - Experimental validation on image datasets.
- Conclusion and future work

Data with low intrinsic dimension: Natural Images[1]

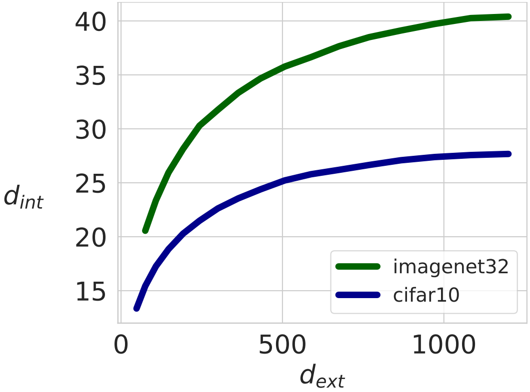
Topographical Ordering of
ImageNet patches

[1] Thiry, Arbel, Belilovsky, and Oyallon, “The Unreasonable Effectiveness of Patches in Deep Convolutional Kernels Methods”.

Data with low intrinsic dimension: Natural Images[1]

Topographical Ordering of ImageNet patches

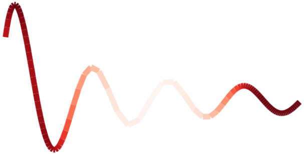
Nearest Neighbor dimension

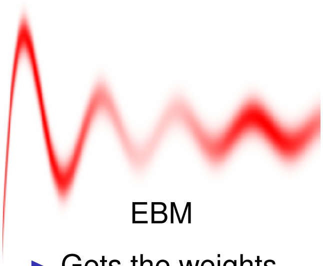


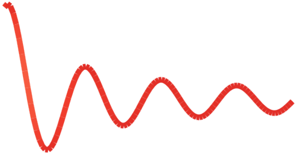
[1] Thiry, Arbel, Belilovsky, and Oyallon, "The Unreasonable Effectiveness of Patches in Deep Convolutional Kernels Methods".

Data with low intrinsic dimension: A toy example

Data with low intrinsic dimension: A toy example

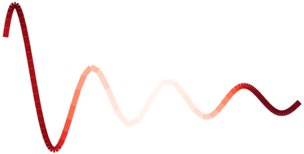

Data


EBM


GAN

- Gets the weights...
- But blurs the samples
- Needs powerful energy models

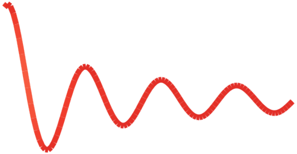
Data with low intrinsic dimension: A toy example



Data

EBM

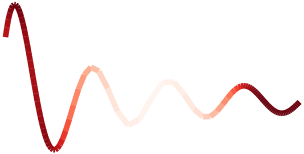
- Gets the weights...
- But blurs the samples
- Needs powerful energy models



GAN

- Gets the support...
- Requires powerful generators
- Wasteful: throws away the critic

Data with low intrinsic dimension: A toy example

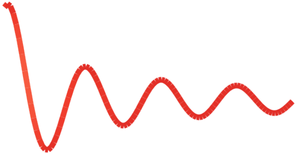


Data

EBM

- Gets the weights...
- But blurs the samples
- Needs powerful energy models

Can we do better?



GAN

- Gets the support...
- Requires powerful generators
- Wasteful: throws away the critic

Generalized Energy-Based Models

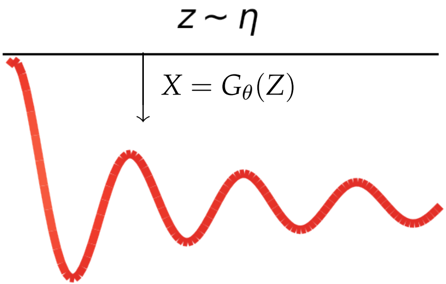
GEBMs are defined by a combination of the two components: *energy* and *base*

Generalized Energy-Based Models

GEBMs are defined by a combination of the two components: *energy* and *base*

- The base learns the low-dimensional support of the data:

$$X \sim \mathbb{B}, \quad \Longleftrightarrow \quad X = G_\theta(Z), \quad Z \sim \eta$$



Generalized Energy-Based Models

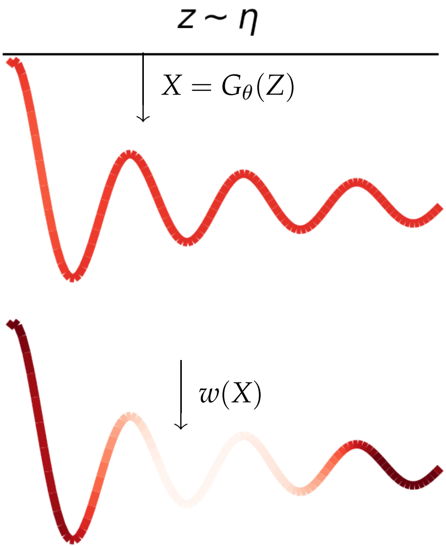
GEBMs are defined by a combination of the two components: *energy* and *base*

- The base learns the low-dimensional support of the data:

$$X \sim \mathbb{B}, \quad \iff \quad X = G_\theta(Z), \quad Z \sim \eta$$

- Samples are re-weighted according to importance weights defined by the energy:

$$w(X) \propto \exp(-E(X))$$

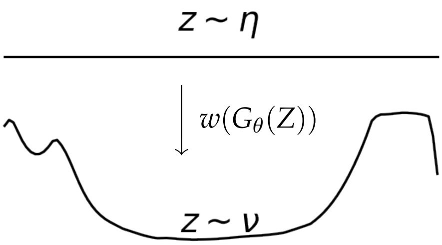


Generalized Energy-Based Models: Latent space view

GEBMs are also obtained by **first re-weighting** the latent then applying G_θ

Generalized Energy-Based Models: Latent space view

GEBMs are also obtained by **first re-weighting** the latent then applying G_θ

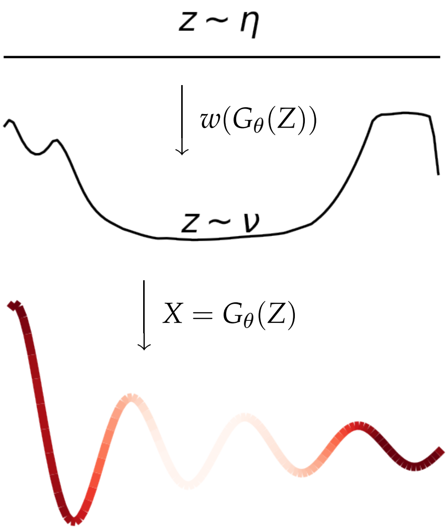


- Latents are sampled according to a 'posterior' distribution:

$$\nu(Z) = \eta(Z) w(G_\theta(Z))$$

Generalized Energy-Based Models: Latent space view

GEBMs are also obtained by **first re-weighting** the latent then applying G_θ



- Latents are sampled according to a 'posterior' distribution:

$$\nu(Z) = \eta(Z) w(G_\theta(Z))$$

- Latents are mapped to sample space using the implicit map G_θ:

$$X = G_\theta(Z)$$

Generalized Energy-Based Models: Why Generalized ?

- A GEBM can be written formally in terms of the *base* $\mathbb{B}_\theta$ and *energy* E:

$$\mathrm{d}\mathbb{Q}(X) \propto \exp(-E(X))\,\mathrm{d}\mathbb{B}_\theta(X)$$

Generalized Energy-Based Models: Why Generalized ?

- A GEBM can be written formally in terms of the *base* $\mathbb{B}_\theta$ and *energy* E:

$$\mathrm{d}\mathbb{Q}(X) \propto \exp(-E(X))\,\mathrm{d}\mathbb{B}_\theta(X)$$

- If the energy E is constant, Q is simply an implicit model:

$$\mathrm{d}\mathbb{Q}(X) = \mathrm{d}\mathbb{B}_\theta(X)$$

Generalized Energy-Based Models: Why Generalized ?

- A GEBM can be written formally in terms of the *base* $\mathbb{B}_\theta$ and *energy* E:

$$\mathrm{d}\mathbb{Q}(X) \propto \exp(-E(X))\,\mathrm{d}\mathbb{B}_\theta(X)$$

- If the energy E is constant, Q is simply an implicit model:

$$\mathrm{d}\mathbb{Q}(X) = \mathrm{d}\mathbb{B}_\theta(X)$$

- If the base is full dimensional and has a density p_θ, Q is a standard EBM:

$$\mathrm{d}\mathbb{Q}(X) \propto \exp(-E(X))p_\theta(X)\,\mathrm{d}X.$$

Generalized Energy-Based Models: Why Generalized ?

- A GEBM can be written formally in terms of the *base* $\mathbb{B}_\theta$ and *energy* E:

$$\mathrm{d}\mathbb{Q}(X) \propto \exp(-E(X))\,\mathrm{d}\mathbb{B}_\theta(X)$$

- If the energy E is constant, Q is simply an implicit model:

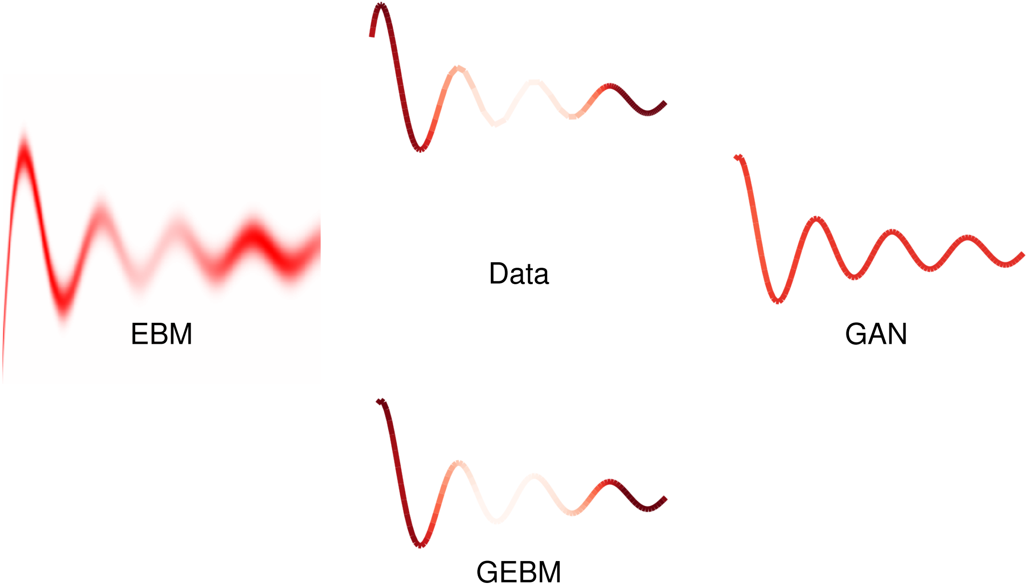
$$\mathrm{d}\mathbb{Q}(X) = \mathrm{d}\mathbb{B}_\theta(X)$$

- If the base is full dimensional and has a density p_θ, Q is a standard EBM:

$$\mathrm{d}\mathbb{Q}(X) \propto \exp(-E(X))p_\theta(X)\,\mathrm{d}X.$$

- GEBM is a generalization of those models that takes the best of both worlds.

GEBMs: The best of both worlds

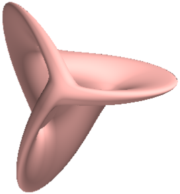


Outline

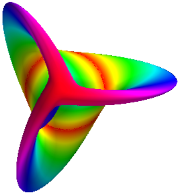
- Data with low intrinsic dimension: The need for new models
- Generalized Energy-Based models: A model with two components
  - The base
  - The energy
- **Training GEBMs: a two stages method**
  - Learning the energy: Generalized Maximum Likelihood Estimation
  - Learning the base : $KALE$ minimization
- Sampling from GEBMs
  - Latent space MCMC
  - Experimental validation on image datasets.
- Conclusion and future work

Training GEBM: A two steps approach

Training the energy: Generalized Maximum Likelihood

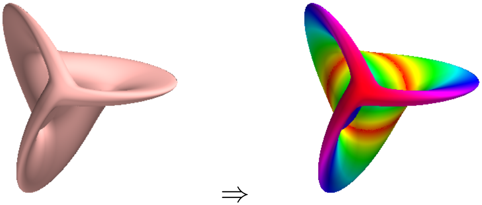


$\Rightarrow$

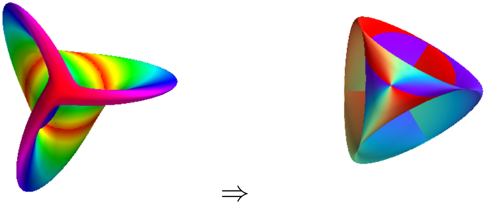


Training GEBM: A two steps approach

Training the energy: Generalized Maximum Likelihood



Training the base: f-divergence minimization (KALE)



Training the energy: Generalized Maximum Likelihood

Definition (Generalized Likelihood)

The expected $\mathbb{B}_\theta$-log-likelihood under a target distribution $\mathbb{P}$ of a GEBM model $\mathbb{Q}$ with base $\mathbb{B}_\theta$ and energy E is defined as

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\int E(x)d\mathbb{P}(x) - \log(Z_{\theta,E}).$$

Training the energy: Generalized Maximum Likelihood

Definition (Generalized Likelihood)

The expected $\mathbb{B}_\theta$-log-likelihood under a target distribution $\mathbb{P}$ of a GEBM model $\mathbb{Q}$ with base $\mathbb{B}_\theta$ and energy E is defined as

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\int E(x) d\mathbb{P}(x) - \log(Z_{\theta,E}).$$

- Dependence on $\mathbb{B}_\theta$ through $Z_{\theta,E} = \mathbb{E}_{\mathbb{B}_\theta}[\exp(-E(X))]$.
- When $KL(\mathbb{P}, \mathbb{B}_\theta)$ is well defined: called Donsker-Varadhan lower bound on KL.
  - Tight when $E(X) = -\log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{B}}(X)\right)$

Training the energy: Generalized Maximum Likelihood

Definition (Generalized Likelihood)

The expected $\mathbb{B}_\theta$-log-likelihood under a target distribution $\mathbb{P}$ of a GEBM model $\mathbb{Q}$ with base $\mathbb{B}_\theta$ and energy E is defined as

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\int E(x)d\mathbb{P}(x) - \log(Z_{\theta,E}).$$

- Dependence on $\mathbb{B}_\theta$ through $Z_{\theta,E} = \mathbb{E}_{\mathbb{B}_\theta}[\exp(-E(X))]$.
- When $KL(\mathbb{P}, \mathbb{B}_\theta)$ is well defined: called Donsker-Varadhan lower bound on KL.
  - Tight when $E(X) = -\log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{B}}(X)\right)$
- However, *Generalized Log-Likelihood* is still well defined when $\mathbb{P}$ and $\mathbb{B}_\theta$ are mutually singular

Training the energy: Generalized Maximum Likelihood

Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E}).$$

- Learn parameters of E using SGD.

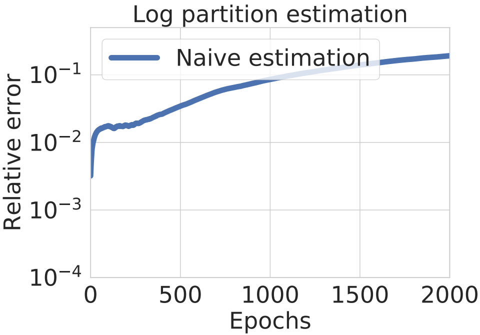
Training the energy: Generalized Maximum Likelihood

Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E}).$$

- Learn parameters of E using SGD.
- Naive estimation of the normalizing constant can have large variance

$$\widehat{\log(Z_{\theta,E})} = \log\left(\frac{1}{N}\sum_{i=1}^{N} exp(-E(X_i))\right)$$



Training the energy: Generalized Maximum Likelihood

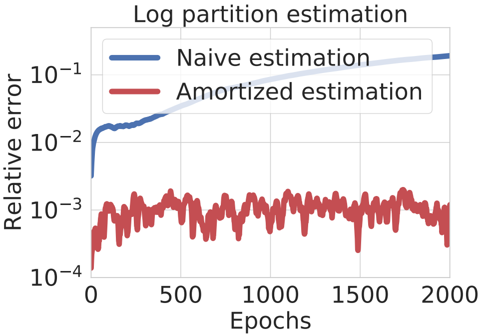
Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E}).$$

- Learn parameters of E using SGD.
- Naive estimation of the normalizing constant can have large variance

$$\widehat{\log(Z_{\theta,E})} = \log\left(\frac{1}{N}\sum_{i=1}^{N} exp(-E(X_i))\right)$$

- Amortized estimation: A better alternative.



Training the energy: Amortized estimation

Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E})$$

- Amortized estimation using a lower-bound on the log-likelihood:

$$\begin{aligned}\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) \geq & -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1 \\ := & \mathcal{F}_{\mathbb{P},\mathbb{B}}(E+c)\end{aligned}$$

Training the energy: Amortized estimation

Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E})$$

- Amortized estimation using a lower-bound on the log-likelihood:

$$\begin{aligned}\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) \geq & -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1 \\ := & \mathcal{F}_{\mathbb{P},\mathbb{B}}(E+c)\end{aligned}$$

- Tight whenever $c = \log(Z_{\theta,E})$

Training the energy: Amortized estimation

Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E})$$

- Amortized estimation using a lower-bound on the log-likelihood:

$$\begin{aligned}\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) \geq & -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1 \\ := & \mathcal{F}_{\mathbb{P},\mathbb{B}}(E+c)\end{aligned}$$

- Tight whenever $c = \log(Z_{\theta,E})$
- Jointly maximizing $\mathcal{F}_{\mathbb{P},\mathbb{B}}(E,c)$ yields the maximum likelihood energy $\mathbb{E}^\star$ and corresponding $c^\star = \log(Z_{\theta,E^\star})$.

Training the energy: Amortized estimation

Learn the energy E using Generalized Log-Likelihood and keep the base $\mathbb{B}_\theta$ fixed.

$$\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) := -\mathbb{E}_{\mathbb{P}}[E(X)] - \log(Z_{\theta,E})$$

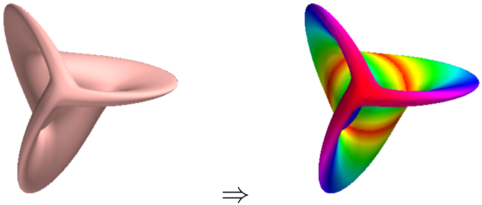
- Amortized estimation using a lower-bound on the log-likelihood:

$$\begin{aligned}\mathcal{L}_{\mathbb{P},\mathbb{B}}(E) \geq & -\mathbb{E}_{\mathbb{P}}[E(X) + c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X) + c))] + 1 \\ := & \mathcal{F}_{\mathbb{P},\mathbb{B}}(E + c)\end{aligned}$$

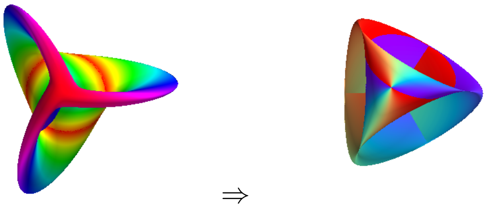
- Tight whenever $c = \log(Z_{\theta,E})$
- Jointly maximizing $\mathcal{F}_{\mathbb{P},\mathbb{B}}(E, c)$ yields the maximum likelihood energy $\mathbb{E}^\star$ and corresponding $c^\star = \log(Z_{\theta,E^\star})$.
- Parameter c keeps a memory of previous mini-batches.

Training GEBM: A two steps approach

Training the energy: Generalized Maximum Likelihood



Training the base: f-divergence minimization (KALE)



Training the base: KALE minimization

- Recall: Optimal energy $E^{\star}$ learned by keeping the base $\mathbb{B}_{\theta}$ fixed and maximizing:

$$\mathcal{F}_{\mathbb{P},\mathbb{B}_{\theta}}(E+c) = -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_{\theta}}[\exp(-(E(X)+c))] + 1$$

[2] Chu, Minami, and Fukumizu, "Smoothness and Stability in GANs".

Training the base: KALE minimization

- Recall: Optimal energy $E^\star$ learned by keeping the base $\mathbb{B}_\theta$ fixed and maximizing:

$$\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E+c) = -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1$$

- Define the KL Approximate Lower-bound Estimator (KALE) to be

$$KALE(\mathbb{P},\mathbb{B}_\theta) := \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E^\star + c^\star)$$

[2] Chu, Minami, and Fukumizu, "Smoothness and Stability in GANs".

Training the base: KALE minimization

- Recall: Optimal energy $E^\star$ learned by keeping the base $\mathbb{B}_\theta$ fixed and maximizing:

$$\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E+c) = -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1$$

- Define the KL Approximate Lower-bound Estimator (KALE) to be

$$KALE(\mathbb{P},\mathbb{B}_\theta) := \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E^\star + c^\star)$$

- KALE defines a divergence between distributions ... if the set of energies $\mathcal{E}$ is rich enough: (ex: an MLP, an RKHS, etc).

[2] Chu, Minami, and Fukumizu, "Smoothness and Stability in GANs".

Training the base: KALE minimization

- Recall: Optimal energy $E^\star$ learned by keeping the base $\mathbb{B}_\theta$ fixed and maximizing:

$$\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E+c) = -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1$$

- Define the KL Approximate Lower-bound Estimator (KALE) to be

$$KALE(\mathbb{P},\mathbb{B}_\theta) := \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E^\star + c^\star)$$

- KALE defines a divergence between distributions ... if the set of energies $\mathcal{E}$ is rich enough: (ex: an MLP, an RKHS, etc).
- Learn the base $\mathbb{B}_\theta$ by minimizing $KALE(\mathbb{P},\mathbb{B}_\theta)$ using SGD.

[2] Chu, Minami, and Fukumizu, "Smoothness and Stability in GANs".

Training the base: KALE minimization

- Recall: Optimal energy $E^\star$ learned by keeping the base $\mathbb{B}_\theta$ fixed and maximizing:

$$\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E + c) = -\mathbb{E}_{\mathbb{P}}[E(X) + c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X) + c))] + 1$$

- Define the KL Approximate Lower-bound Estimator (KALE) to be

$$KALE(\mathbb{P}, \mathbb{B}_\theta) := \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E^\star + c^\star)$$

- KALE defines a divergence between distributions ... if the set of energies $\mathcal{E}$ is rich enough: (ex: an MLP, an RKHS, etc).
- Learn the base $\mathbb{B}_\theta$ by minimizing $KALE(\mathbb{P}, \mathbb{B}_\theta)$ using SGD.
- Is the gradient well-defined? Is it smooth enough?

[2] Chu, Minami, and Fukumizu, "Smoothness and Stability in GANs".

Training the base: KALE minimization

- Recall: Optimal energy $E^{\star}$ learned by keeping the base $\mathbb{B}_\theta$ fixed and maximizing:

$$\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E+c) = -\mathbb{E}_{\mathbb{P}}[E(X)+c] - \mathbb{E}_{\mathbb{B}_\theta}[\exp(-(E(X)+c))] + 1$$

- Define the KL Approximate Lower-bound Estimator (KALE) to be

$$KALE(\mathbb{P},\mathbb{B}_\theta) := \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E^{\star}+c^{\star})$$

- KALE defines a divergence between distributions ... if the set of energies $\mathcal{E}$ is rich enough: (ex: an MLP, an RKHS, etc).
- Learn the base $\mathbb{B}_\theta$ by minimizing $KALE(\mathbb{P},\mathbb{B}_\theta)$ using SGD.
- Is the gradient well-defined? Is it smooth enough?
- Lack of smoothness can result in instabilities during training[2]

[2] Chu, Minami, and Fukumizu, "Smoothness and Stability in GANs".

Training the base: Smoothness of KALE

- The loss results from an optimization:

$$KALE(\mathbb{P}, \mathbb{B}_\theta) = \sup_{E,c} \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E + c)$$

- The gradient is expected to be of the form:

$$\nabla_\theta KALE(\mathbb{P}, \mathbb{B}_\theta) = \nabla_\theta \mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E^\star + c^\star)$$

- No guarantees this holds in general: needs additional assumptions.
- Typical assumptions rely on convexity[3] of $\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E + c)$ in the parameters of E , or measure smoothness assumptions[4] : too strong in this case.

---

[3] Sanjabi, Ba, Razaviyayn, and Lee, “Solving Approximate Wasserstein GANs to Stationarity”.
[4] Chu, Minami, and Fukumizu, “Smoothness and Stability in GANs”.

Training the base: Smoothness of KALE

Theorem (An enveloppe theorem)

$KALE(\mathbb{P}, \mathbb{B}_\theta)$ is Lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla_\theta KALE(\mathbb{P}, \mathbb{B}_\theta) = \mathbb{E}_{\nu_{\theta,E^\star}}[\nabla_x E^\star(G_\theta(Z))\nabla_\theta G_\theta(Z)]$$

with $\nu_{\theta,E^\star}$ being the re-weighted latent distribution: $\nu_{\theta,E^\star}(Z) \propto \exp(-E^\star(G_\theta(Z)))$.

Training the base: Smoothness of KALE

Theorem (An enveloppe theorem)

$KALE(\mathbb{P}, \mathbb{B}_\theta)$ *is Lipschitz and differentiable for almost all* $\theta \in \Theta$ *with:*

$$\nabla_\theta KALE(\mathbb{P}, \mathbb{B}_\theta) = \mathbb{E}_{\nu_{\theta,E^\star}}[\nabla_x E^\star(G_\theta(Z))\nabla_\theta G_\theta(Z)]$$

with $\nu_{\theta,E^\star}$ *being the re-weighted latent distribution:* $\nu_{\theta,E^\star}(Z) \propto \exp(-E^\star(G_\theta(Z)))$.

Assumptions:

- Energies in $\mathcal{E}$ parameterized by $\psi \in \Psi$, where Ψ is compact. Jointly continuous in (ψ, x) and L-smooth w.r.t. x.
- $(\theta, z) \mapsto G_\theta(z)$ L-Lipschitz in z and smooth wrt θ.

Training the base: Smoothness of KALE

Theorem (An enveloppe theorem)

$KALE(\mathbb{P}, \mathbb{B}_\theta)$ *is Lipschitz and differentiable for almost all* $\theta \in \Theta$ *with:*

$$\nabla_\theta KALE(\mathbb{P}, \mathbb{B}_\theta) = \mathbb{E}_{\nu_{\theta,E^\star}}[\nabla_x E^\star(G_\theta(Z))\nabla_\theta G_\theta(Z)]$$

with $\nu_{\theta,E^\star}$ *being the re-weighted latent distribution:* $\nu_{\theta,E^\star}(Z) \propto \exp(-E^\star(G_\theta(Z)))$.

Assumptions:

- Energies in $\mathcal{E}$ parameterized by $\psi \in \Psi$, where Ψ is compact. Jointly continuous in (ψ, x) and L-smooth w.r.t. x.
- $(\theta, z) \mapsto G_\theta(z)$ L-Lipschitz in z and smooth wrt θ.

Proof idea:

- Characterization of differentiability for supremum-type functions[5]:
  - Expressions for left and right partial derivatives of the loss. Expressions match when $\theta \mapsto E^\star_\theta$ is continuous.
  - Differentiability holds iff $\theta \mapsto E^\star_\theta$ is continuous.

[5] Milgrom and Segal, “Envelope Theorems for Arbitrary Choice Sets”.

Training the base: Smoothness of KALE

Theorem (An enveloppe theorem)

$KALE(\mathbb{P}, \mathbb{B}_\theta)$ is Lipschitz and differentiable for almost all $\theta \in \Theta$ with:

$$\nabla_\theta KALE(\mathbb{P}, \mathbb{B}_\theta) = \mathbb{E}_{\nu_{\theta,E^\star}}[\nabla_x E^\star(G_\theta(Z))\nabla_\theta G_\theta(Z)]$$

with $\nu_{\theta,E^\star}$ being the re-weighted latent distribution: $\nu_{\theta,E^\star}(Z) \propto \exp(-E^\star(G_\theta(Z)))$.

Assumptions:

- Energies in $\mathcal{E}$ parameterized by $\psi \in \Psi$, where Ψ is compact. Jointly continuous in (ψ, x) and L-smooth w.r.t. x.
- $(\theta, z) \mapsto G_\theta(z)$ L-Lipschitz in z and smooth wrt θ.

Proof idea:

- Characterization of differentiability for supremum-type functions[5]:
  - Expressions for left and right partial derivatives of the loss. Expressions match when $\theta \mapsto E_\theta^\star$ is continuous.
  - Differentiability holds iff $\theta \mapsto E_\theta^\star$ is continuous.
- Prove differentiability using Radamacher theorem.

[5] Milgrom and Segal, “Envelope Theorems for Arbitrary Choice Sets”.

Training GEBM: Summary

GEBMs are defined by:

$$d\mathbb{Q}_{\theta,E}(X) \propto exp(-E(X))\,\mathrm{d}\mathbb{B}_{\theta}(X)$$

Training GEBM: Summary

GEBMs are defined by:

$$d\mathbb{Q}_{\theta,E}(X) \propto exp(-E(X))\,\mathrm{d}\mathbb{B}_{\theta}(X)$$

Training alternates between:

- Training the energy: Maximize the lower-bound $\mathcal{F}_{\mathbb{P},\mathbb{B}_{\theta}}(E+c)$ on the generalized log-likelihood.
- Training the base: Minimize $KALE(\mathbb{P},\mathbb{B}_{\theta})$

Training GEBM: Summary

GEBMs are defined by:

$$d\mathbb{Q}_{\theta,E}(X) \propto exp(-E(X))\,\mathrm{d}\mathbb{B}_\theta(X)$$

Training alternates between:

- Training the energy: Maximize the lower-bound $\mathcal{F}_{\mathbb{P},\mathbb{B}_\theta}(E+c)$ on the generalized log-likelihood.
- Training the base: Minimize $KALE(\mathbb{P},\mathbb{B}_\theta)$

Can we guarantee that the GEBM $\mathbb{Q}$ is getting closer to $\mathbb{P}$?

Training GEBM: Summary

GEBMs are defined by:

$$d\mathbb{Q}_{\theta,E}(X) \propto exp(-E(X))\, \mathrm{d}\mathbb{B}_{\theta}(X)$$

Training alternates between:

- Training the energy: Maximize the lower-bound $\mathcal{F}_{\mathbb{P},\mathbb{B}_{\theta}}(E+c)$ on the generalized log-likelihood.
- Training the base: Minimize $KALE(\mathbb{P}, \mathbb{B}_{\theta})$

Can we guarantee that the GEBM $\mathbb{Q}$ is getting closer to $\mathbb{P}$?

Theorem

If the set of energies $\mathcal{E}$ is convex, then:

$$KALE(\mathbb{P}, \mathbb{Q}_{\theta,E^{\star}}) \leq 2KALE(\mathbb{P}, \mathbb{B}_{\theta})$$

where $E^{\star}$ maximizes the generalized $\mathbb{B}_{\theta}$ log-likelihood

Training GEBM: Does it really learn Maximum likelihood ?

Particular instance for GEBM:

- The base $\mathbb{B}_\theta(X)$ is a Real NVP[6] (closed form density $exp(h_\theta(X))$)
- The Energy is of the form $E(X) = r_\psi(X) - h_\theta(X)$
- For this choice, GEBM is equivalent to an EBM of the form

$$d\mathbb{Q}_{\theta,E}(X) \propto \exp(-r_\psi(X))\,\mathrm{d}X.$$

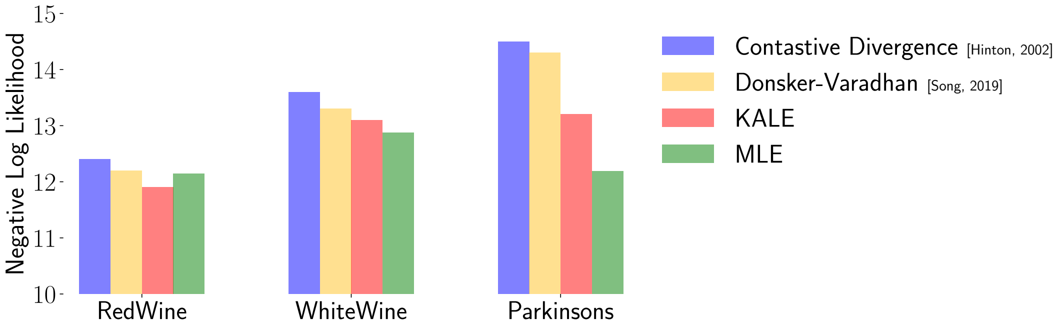
[6] Dinh, Sohl-Dickstein, and Bengio, "Density estimation using Real NVP".

Training GEBM: Does it really learn Maximum likelihood ?

Particular instance for GEBM:

- The base $\mathbb{B}_\theta(X)$ is a Real NVP[6] (closed form density $exp(h_\theta(X))$)
- The Energy is of the form $E(X) = r_\psi(X) - h_\theta(X)$
- For this choice, GEBM is equivalent to an EBM of the form

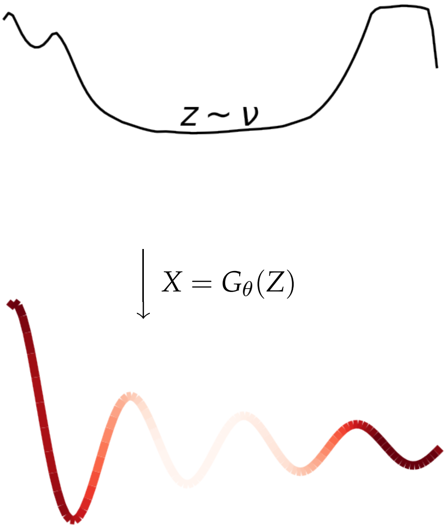
$$d\mathbb{Q}_{\theta,E}(X) \propto \exp(-r_\psi(X))\, \mathrm{d}X.$$



[6] Dinh, Sohl-Dickstein, and Bengio, "Density estimation using Real NVP".

Sampling from GEBMs: Latent space MCMC

GEBMs are defined by $d\mathbb{Q}_{\theta,E}(X) = w(X)\,\mathrm{d}\mathbb{B}_\theta(X)$ with $w(X) \propto exp(-E(X))$.



- Latents are sampled according to a 'posterior' distribution:

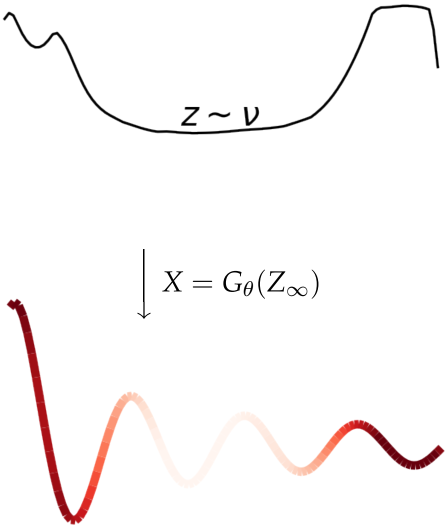
$$\nu(Z) = \eta(Z) w(G_\theta(Z))$$

- Latents are mapped to sample space using the implicit map G_θ:

$$X = G_\theta(Z)$$

Sampling from GEBMs: Latent space MCMC

GEBMs are defined by $d\mathbb{Q}_{\theta,E}(X) = w(X)\, \mathrm{d}\mathbb{B}_\theta(X)$ with $w(X) \propto exp(-E(X))$.



- Latents are sampled according to a 'posterior' distribution:

$$\nu(Z) = \eta(Z) w(G_\theta(Z))$$

- In practice, use MCMC

$$W_{k+1} \sim \mathcal{N}(0, I)$$
$$Z_{k+1} = Z_k + \gamma \nabla_z \log \nu(Z_k) + \sqrt{2\gamma} W_{k+1}$$

- Latents are mapped to sample space using the implicit map G_θ:

$$X = G_\theta(Z)$$

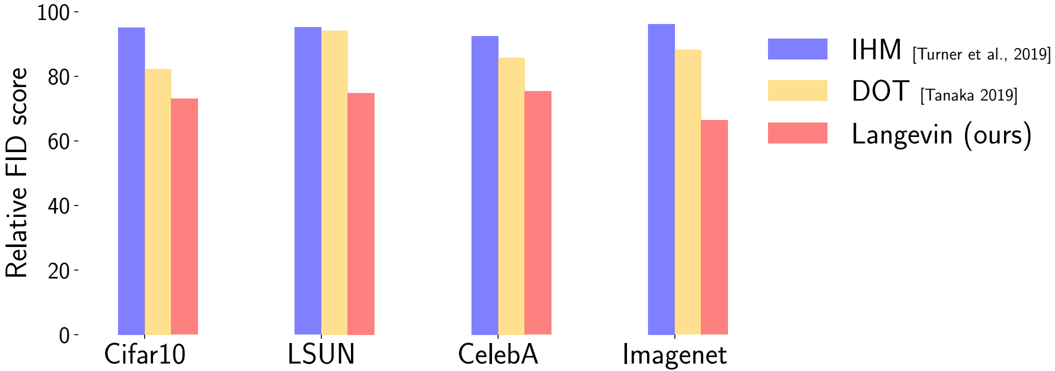
Outline

- Data with low intrinsic dimension: The need for new models
- Generalized Energy-Based models: A model with two components
  - The base
  - The energy
- Training GEBMs: a two stages method
  - Learning the energy: Generalized Maximum Likelihood Estimation
  - Learning the base : $KALE$ minimization
- **Sampling from GEBMs**
  - Latent space MCMC
  - Experimental validation on image datasets.
- Conclusion and future work

Sampling from GEBMs: Latent space MCMC

Sampling for Generalized EBMs

- Relative FID score: $\frac{FID(\mathbb{Q}_{\theta,E})}{FID(\mathbb{B}_{\theta})}$.



For a given base $\mathbb{B}_\theta$ and energy E trained using KALE, samples from the GEBM are always better (FID score) than samples from the base alone.

Sampling from GEBMs: Jumping between modes

Other samplers (ex. Hamiltonian Monte Carlo) allows better mode exploration

Summary

- GEBMs are models tailored for data with low intrinsic dimension
- Combine the strength of both Implicit (the base) and Explicit models (the energy)
- Two stages training : alternating optimization on the base and energy
- Sampling performed by Latent space MCMC
- Improves over sampling from the base alone (as done in GANs)

Summary

- GEBMs are models tailored for data with low intrinsic dimension
- Combine the strength of both Implicit (the base) and Explicit models (the energy)
- Two stages training : alternating optimization on the base and energy
- Sampling performed by Latent space MCMC
- Improves over sampling from the base alone (as done in GANs)

Future directions:

- Can training GEBMs be improved?
  - Better than a two-step training (one step?)
  - Is latent space MCMC beneficial during training[7]?
- Generalization of GEBMs
  - Do the modes defined by the energy match training samples? Is it bad[8]?

[7] Wu et al., “LOGAN: Latent Optimisation for Generative Adversarial Networks”.
[8] Belkin, Rakhlin, and Tsybakov, “Does data interpolation contradict statistical optimality?”

Thank you!

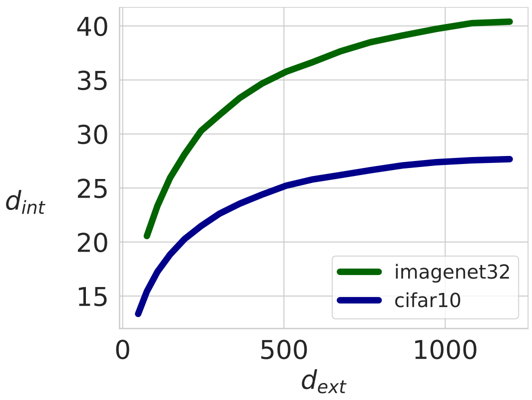
Estimating Intrinsic dimension[9]

- For a sample X, find the k-NNs $X_1, \dots, X_k$
- Compute distances $T_j(X) = \|X - X_j\|$
- Estimate dimension at point X:

$$d(X) = \left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(X)}{T_j(X)}\right]^{-1}$$

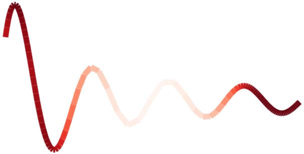
- Average over several points X and values of k.

Nearest Neighbor dimension



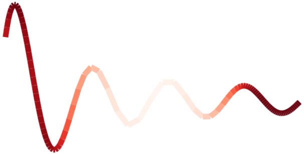
[9] Levina and Bickel, “Maximum likelihood estimation of intrinsic dimension”.

Data with low intrinsic dimension: A toy example



Data

Data with low intrinsic dimension: A toy example



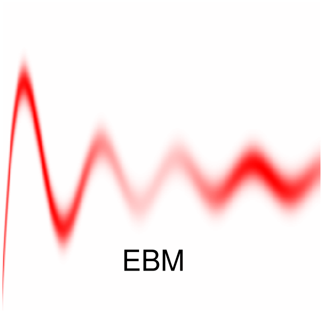
Data

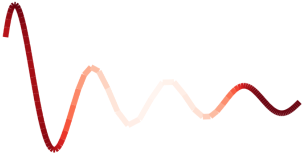
$$z \sim \mathit{Unif}[0,1]$$

$$\widetilde{z} = \overset{\downarrow}{\tau}(z)$$

$$X = \overset{\downarrow}{G}_{\theta^\star}(\widetilde{z}), \quad X_1 = \widetilde{z}$$

Data with low intrinsic dimension: A toy example





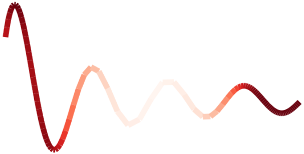
Data

$$z \sim Unif[0,1]$$

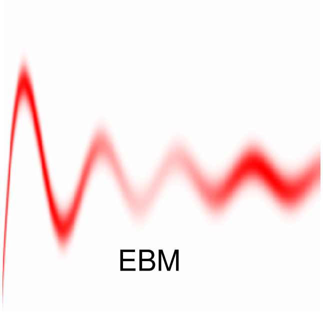
$$\tilde{z} = \overset{\downarrow}{\tau}(z)$$

$$X = \overset{\downarrow}{G}_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

Data with low intrinsic dimension: A toy example



Data



$$p(X) \propto \exp(-E(X))$$

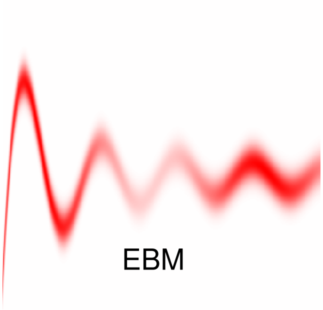
$$E(X) = \frac{1}{2\sigma^2} \| G_\theta(X_1) - X \|^2 + A_\theta(X_1)$$

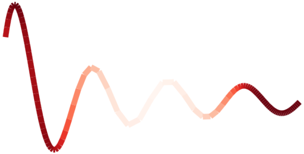
$$z \sim Unif[0,1]$$

$$\tilde{z} = \overset{\downarrow}{\tau}(z)$$

$$X = \overset{\downarrow}{G}_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

Data with low intrinsic dimension: A toy example





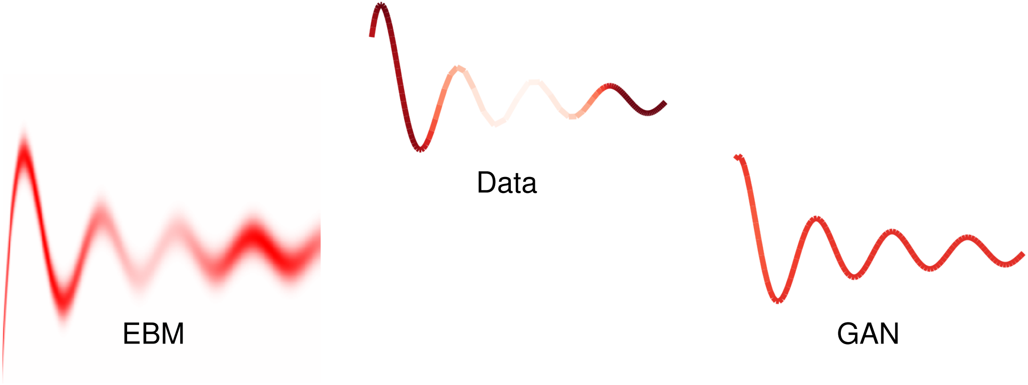
Data

$$z \sim \mathit{Unif}[0,1]$$

$$\tilde{z} = \overset{\downarrow}{\tau}(z)$$

$$X = \overset{\downarrow}{G}_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

Data with low intrinsic dimension: A toy example

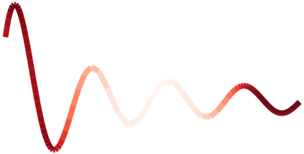


$$z \sim Unif[0,1]$$

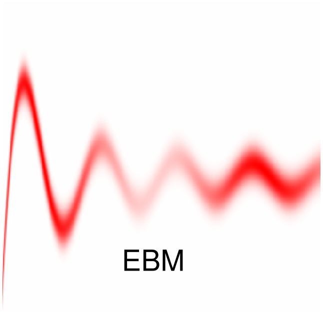
$$\tilde{z} = \overset{\downarrow}{\tau}(z)$$

$$X = \overset{\downarrow}{G}_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

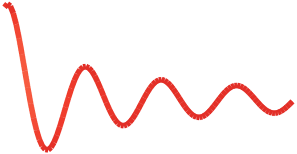
Data with low intrinsic dimension: A toy example



Data



EBM



GAN

$$z \sim Unif[0, 1]$$

$$\tilde{z} = \overset{\downarrow}{\tau}(z)$$

$$X = \overset{\downarrow}{G}_{\theta^\star}(\tilde{z}), \quad X_1 = \tilde{z}$$

Generator

$$z \sim unif[0, 1]$$

$$X = \overset{\downarrow}{G}_{\theta}(z)$$

Critic

$$MLP(X)$$

Belkin, Mikhail, Alexander Rakhlin, and Alexandre B Tsybakov. "Does data interpolation contradict statistical optimality?" In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR. 2019, pp. 1611–1619.

Chu, Casey, Kentaro Minami, and Kenji Fukumizu. "Smoothness and Stability in GANs". In: 2019.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: (May 2016). eprint: 1605.08803. URL: https://arxiv.org/abs/1605.08803.

Levina, Elizaveta and Peter J. Bickel. "Maximum likelihood estimation of intrinsic dimension". In: Advances in neural information processing systems 17. MIT Press, 2004.

Milgrom, Paul and Ilya Segal. "Envelope Theorems for Arbitrary Choice Sets". In: Econometrica 70.2 (2002).

Sanjabi, Maziar et al. "Solving Approximate Wasserstein GANs to Stationarity". In: arXiv:1802.08249 [cs, math, stat] (Feb. 2018). arXiv: 1802.08249. URL: http://arxiv.org/abs/1802.08249 (visited on 04/07/2018).

Thiry, Louis et al. "The Unreasonable Effectiveness of Patches in Deep